91602
new 82-10

COMPLIMENTARY COPY
Please Credit Ontario Hydro Research if Reproduced in Whole or in Part

ontario hydro research division

THE IDENTIFICATION OF HIGH SO_2 EPISODES AT NANTICOKE GS FROM METEOROLOGICAL AND LOAD DATA

Report No 82-509-K

T.J. Barker
Mathematician - Reliability and Statistics Section
Operations Research Department

ABSTRACT

This report is an extension of the preliminary statistical analysis of the Nanticoke (1976-78) database described in an earlier Research Division Report/1/. The preliminary analysis identified those meteorological conditions in the database which are associated with high and low hourly monitored SO_2 values in the vicinity of Nanticoke GS. This work identifies those meteorological scenarios which are associated exclusively with high monitored SO_2 values, as distinct from all other situations. This allows such episodes to be readily identified on the basis of a small number of basic meteorological parameters. These meteorological scenarios are discussed and their wider applicability assessed using data from Lakeview GS involving only a small number of days.

job	file	date	report no.
740635-646-091	809.710.01	December 16, 1982	82-509-K

91602
new 82-10

ontario hydro
research division

EXECUTIVE SUMMARY

THE IDENTIFICATION OF HIGH SO_2 EPISODES AT NANTICOKE GS FROM METEOROLOGICAL AND LOAD DATA

T.J. Barker
Mathematician - Reliability and Statistics Section
Operations Research Department

Meteorological and station data from Nanticoke GS involving 80 days between March 1976 to January 1978 were analyzed in a previous report (Research Division Report No 82-56-K). The analysis identified a set of meteorological variables which could be used to discriminate between those hours where the SO_2 readings exceeded 89 ppb (high hours) and those where it was less than 17 ppb (low hours).

The present study, instead of considering high hours, examined the data on the basis of high days ($\geq$120 ppb) and found a statistically significant separation between those days and all other days. The variables used in this separation were total generation (MW), wind speed at ground (WSG) and wind direction shear (WDS) together with a time requirement, namely that the conditions prevail for at least two observations (1-2 h). The scenarios identified for high SO_2 cases at Nanticoke are as shown in the following Table.

NANTICOKE HIGH SO_2 (> 120 ppb) Scenarios

MW	WSG (m/sec)	WDS (deg/m)	Remarks
> 1000	$\geq$ 8.5	$\leq$ 0.07	High wind fumigation
> 1900	$\geq$ 3.5 and < 8.5	$\geq$ 0.08 and <0.16	Activity in 200-500 m layer, possibly movement or growth of inversion

Note: The given conditions on MW, WSG, WDS must be satisfied simultaneously in each of the two cases.

job	file	date	report no.
740635-646-091	809.710.01	December 16, 1982	82-509-K

The importance of wind speed in identifying high SO_2 prone days was confirmed by the results of an analysis of 22 days data from Lakeview GS. However, wind direction shear which at Nanticoke appeared to play the part of an index of the movement or growth of inversions and thereby indicate high SO_2 readings, played no such role at Lakeview.

TABLE OF CONTENTS

		PAGE
1.	INTRODUCTION	1
2.	IDENTIFICATION OF SIGNIFICANT METEOROLOGICAL VARIABLES FOR HIGH SO_2 EPISODES AT NANTICOKE	2
	2.1 Variables Significant in Discriminating Between High and Low SO_2 Situations	2
	2.2 Analysis of Subsets of Variables	2
	2.3 The Best Subset	4
3.	NANTICOKE HIGH SO_2 SCENARIOS	4
	3.1 Data	4
	3.2 Criteria for Identifying High SO_2 Days From Meteorological and Load Data	4
	3.3 Predicted and Actual High Days	5
	3.4 Misclassified Days	8
4.	LAKEVIEW HIGH SO_2 SCENARIOS	9
	4.1 Data	9
	4.2 Criteria for Identifying High SO_2 Days From Meteorological and Load Data	9
	4.3 Predicted and Actual High Days	10
	4.4 Misclassified Days	10
5.	COMPARISION BETWEEN THE FINDINGS FOR NANTICOKE AND LAKEVIEW	12
6.	SUMMARY	13
REFERENCES		14
DISTRIBUTION		last page

91601
new 82-10

ontario hydro
research division

To Mr. F.J. Kee
Director of Research

THE IDENTIFICATION OF HIGH SO_2 EPISODES AT NANTICOKE GS FROM METEOROLOGICAL AND LOAD DATA

1. INTRODUCTION

The identification of meteorological conditions associated with high SO_2 values in the vicinity of a generating station is of considerable importance in the validation of physical models of dispersion. At the request of the Chemical Research Department, the Statistics Unit undertook a preliminary statistical analysis of the Nanticoke* database (1976-78), described in an earlier Research Division Report/1/. This analysis succeeded in identifying a subset of variables adequate for discriminating between the highest 10% of hours (over approx 89 ppb) and the lowest 10% of hours (below approx 17 ppb).

The more difficult task of identifying those variables associated exclusively with high SO_2 values at Nanticoke has been attempted in this work. The aim has been to specify the smallest subset of meteorological variables on the basis of which any given set of conditions could be accurately classified as high SO_2 prone or not. In the present project the emphasis changed from considering hourly SO_2 levels to examining the daily SO_2 situation (see footnote in Section 2.2).

This report describes the identification problem and discusses the identified scenarios for Nanticoke. An analysis was also performed on some values for the identified meteorological variables, from 22 days, obtained for Lakeview GS (1979-81) provided by the Chemical Research Department. The results from this small dataset are compared to those obtained for Nanticoke.

* Nanticoke and Lakeview are used throughout to indicate the areas in the immediate vicinity of Nanticoke GS and Lakeview GS.

2. IDENTIFICATION OF SIGNIFICANT VARIABLES FOR HIGH SO_2 EPISODES AT NANTICOKE

2.1 Variables Significant in Discriminating Between High and Low SO_2 Situations

In /1/, the variables having significant ability to discriminate between high and low hours at Nanticoke were identified to be the following:

Wind direction (ground level)	(WD)
Wind speed (ground level)	(WSG)
Dry-bulb temperature	(DT)
Wet-bulb temperature	(WT)
Total megawatts generated	(MW)
Wind direction shear	(WDS)
Wind speed shear	(WSS)
Height of inversion base	(HIB)

Logically any variable set which can be used to separate high SO_2 cases from all others must be a subset of the above set. Therefore, the starting point for the present study was a search of all the subsets of the above variable set to find those which had distinct values for high SO_2 cases.

2.2 Analysis of Subsets of Variables

The data consisted of 469 hourly observations covering 80 days at Nanticoke and consisting of values of the above variables together with the corresponding monitored SO_2 value. This data-set is described in detail in /1/.

A number of statistical pattern recognition techniques were employed in an attempt to spatially separate the high SO_2 cases from all others. The main techniques used were variations on Andrews plots/2/ and the Nearest Neighbour Method/3/. All subsets of the variables were used with both the actual variable values and their rank transforms (see /1/).

These exhaustive procedures proved largely inconclusive with no discernable separation between the two groups being apparent. However, a simply devised 'scoring' technique did exhibit some skill in separating a number of high SO_2 cases (together with a larger number of non-high) from the bulk of cases. These scores were the sum of the individual variable 'coded' values for each particular case. For each individual variable a code was assigned according to that variable's value in relation to the means of that variable calculated for the high and low SO_2 groups of hourly cases (as defined in 2.1). Thus, a variable value which was close (a quarter of the separation (D) between

high and low means) to the high group mean would be assigned +1. Similarly the code would be -1 for a value close to the low mean. Values close to the high group mean whose distance from the low mean exceeded D were coded +2. Similarly values close to the low mean and at a distance greater than D from the high mean were coded -2. Hence, it was hoped that a large positive (negative) total score would indicate a degree of similarity with the high (low) group. Treating each hourly observation as one case, case scores were calculated and plotted against SO_2 for many different subsets of variables. The one subset for which the plot appeared to separate some high SO_2 cases consisted of WT, WSG, HIB. However, the skill of the "classification" based on this subset was still relatively poor, missing nearly one third of the high cases and giving high positive scores to a large number of non-high cases.

The classification skill in terms of the percentage of correctly and incorrectly classified cases improved when each day was treated as one case instead of each hour. The scoring procedure in this new analysis operated on the hour of maximum SO_2 concentration of each day. The scoring technique treated the data as a set of daily records to be classified as "high days" or "non-high days" dependent on whether or not its daily maximum SO_2 value was high* (over approximately 100 ppb). This treatment attempted to recognize the diurnal nature of many of the local meteorological conditions and the importance of the effect of time period in the diffusion of pollutants. When the scoring technique was applied to days a separation between high and non-high cases became apparent. The WT variable was now found to be non-contributory to the separation and could possibly have been acting as a surrogate for time in the hourly analysis. HIB was dropped in favour of WDS which was found to be a more reliable indicator of local "activity" in the 200-500 m range of the atmosphere. An observed downward movement of the HIB often accompanied a high SO_2 episode but not always.

* In the earlier analysis reported in /1/, high SO_2 was defined arbitrarily to be any value which belonged to the highest 10% of the sample of available hourly observations, which made the cut off value for the high observations 89 ppb. When the emphasis in the analysis changed to considering daily scenarios it was found that over 20% of the 80 days in the database had daily maxima of 89 ppb or more. At the initial stage of the daily analysis the definition of high was altered to be over 100 ppb, constituting some 18% of the sample. It was then found that this too gave indistinct groupings of cases and a further refinement then treated high cases as greater than 120 ppb, constituting the top 15% of the sample (see Section 3).

In this way a "best" subset of variables was determined, in the sense that the proximity of their values to critical values (associated with high SO_2 episodes) appeared to indicate whether or not the day in question did contain a high SO_2 episode.

2.3 The Best Subset

The best variable set for indicating high SO_2 conditions was found to be MW, WSG and WDS. By choosing suitable values for these parameters as criteria, good separation between high and non-high cases was achieved. These values appeared to act as threshold or critical values. It was found that occasions where a variable exceeded the critical value by a large amount did not necessarily correspond to higher SO_2 values than those where a critical value was only marginally exceeded. One likely explanation for this is that the meteorological information available was insufficient to characterize the local meteorological system and allow a "clustering" of high SO_2 cases to be detected.

The results of using this best subset of variables in classifying all the available Nanticoke data into high and non-high SO_2 days are discussed in the following section.

3. NANTICOKE HIGH SO_2 SCENARIOS

3.1 Data

After removing observations with missing values for either MW, WSG, WDS or HIB there remained 328 hourly observations on 80 days. Values of the HIB variable were considered worth retaining in the study for their importance as background information on inversion behaviour likely to be required in any "post prediction" analysis of results.

3.2 Criteria for Identifying High SO_2 Days from Meteorological and Load Data

From the results of the scoring technique mentioned in the previous section, it was found that particular values of MW, WSG, WDS were associated with the occurrence of high SO_2 readings. However classification of hourly observations according to particular meteorological values proved inconclusive and only when days were scored was an improvement in the percentage of correct classifications noted. Furthermore it was evident from inspection of the incorrectly classified days that certain other conditions must be met, relating to time. The meteorological system had to pass through either one of two 'states' in a particular span of time in order that a high SO_2 value should occur. The meteorological values

corresponding to these two states were determined empirically in the form of minimum conditions which must be met by at least two hourly observations during any given day.

The first state, that of 'high wind fumigation' may occur if two or more observations during any day satisfy all the following conditions:

MW > 1000
WSG $\geq$ 8.5 m/sec
WDS $\leq$ 0.07 degrees/m

The second state, which perhaps corresponds to the condition known as 'strong vertical mixing' may occur if two or more hourly observations during any day satisfy all the following conditions.

MW > 1900
3.5 $\leq$ WSG < 8.5 m/sec
0.08 $\leq$ WDS < 0.16 degrees/m

The first situation corresponds to a strong wind with constant direction in the 200-500 m range. The low WDS also seems to signify that the inversion base is not moving rapidly up or down.

The second situation corresponds to high generation, moderate wind speed accompanied by "some kind of activity" in the 200-500 m range. Sometimes this signifies that the inversion base is descending (from examination of the HIB values) but not always.

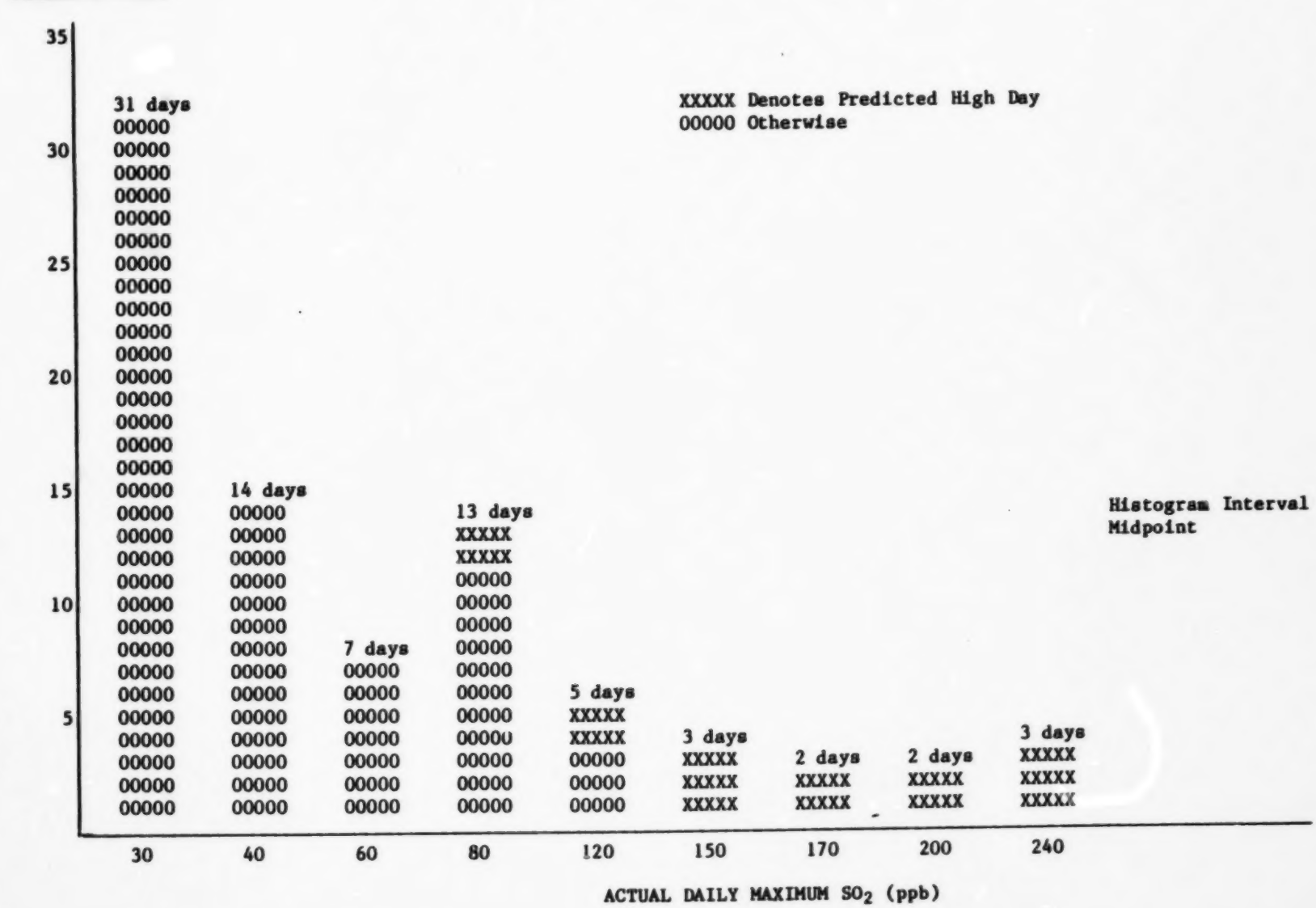
The time stipulation in the criteria attempt to differentiate between a single brief episode of high SO_2 prone conditions and a sufficiently long episode to produce a high monitor reading, which is probably of the order of 1-2 hours.

As a test of the validity of the criteria described, all the days were classified as high or non-high days according to the above criteria. The results of this exercise are discussed in the following section.

3.3 Predicted and Actual High Days

In Figure 1 the predicted high days, those satisfying either of the two sets of empirical criteria, are plotted against the

FIGURE 1: ACTUAL HIGH DAYS AND PREDICTIONS FOR 80 DAYS AT NANTICOKE



actual daily maximum monitored SO_2 values for the 80 days available from Nanticoke. The data are plotted such that where a day was classified as "high" it appears as XXXXX and where it was not as 00000. In all, 14 days were classified as "high" of which 11 were greater than 120 ppb (see note in Section 2.2).

The contingency table for predicted and observed values is therefore of the form given in Table I.

TABLE I: PREDICTED AND ACTUAL HIGH SO_2 CASES

	Number Days > 120 ppb	Number Days $\leq$ 120 ppb
Predicted High Days	11 (2)	3 (0)
Predicted Non-High Days	1 (10)	65 (68)

The numbers in brackets are obtained by considering the numbers in each category that would be obtained by 'fixing' the criteria; that is, choosing two cases as representative and basing the criteria on their respective MW, WSG and WDS values. Obviously the application of two such sets of criteria to the same data would most likely result in correctly identifying at two high cases and, 68 non-high cases; with the remaining high cases misclassified as non-high. Such a fix would be correct 70 times out of 80 and would therefore be considered better than chance. It is possible that such criteria could correctly identify more than two high cases but this would require that two or more high cases have exactly the same MW, WSG and WDS values, which is most unlikely. It is for this reason that the results of the prediction exercise are split into two parts, a test of skill at classifying high cases and a test of skill at classifying non-high cases. The two parts constitute the vertical halves of the above Table I. A Fishers Exact test was performed on these two subtables to establish whether the selected meteorological criteria performed significantly better at classifying daily SO_2 levels than any randomly selected set of two criteria (based on observed high SO_2 cases) would have done.

The null hypothesis in both cases was, therefore, that the criteria were merely set equal to the observed values of two high SO_2 cases.

The probability of obtaining the observed number of correct "high" cases under the null hypothesis was calculated to be less

than 5 in 1000. Not surprisingly, the probability of obtaining the observed number of correct non-high cases under the null hypothesis is much larger; in the given case, it was found to be approximately 12 in 100.

Thus, it can be concluded that the two criteria discussed perform significantly better than "a fix" at identifying high SO_2 meteorological scenarios.

It is worth noting that since the given criteria constitute such a clear cut boundary between high and non-high cases a reversal of the criteria will have very high skill in predicting non-high cases, as can be seen from Figure 1. The great majority of the non-high (below 120 ppb) days are correctly identified as such and denoted by 00000.

A description of those cases which were misclassified by the procedure is instructive since even in these cases the applied criteria did pick out the daily maxima. The four misclassified cases are described below.

3.4 Misclassified Days

Date	Max-SO_2 (ppb)	Remarks
77/2/8	130	Classified as 'low'. The inversion base hovers at approximately 200 m throughout the monitored period. Wind speeds are moderate 6 to 7 m/sec, wind direction shears are 0.03-0.06 deg/m.
76/10/29	99	The identified 'high' hourly observations do correspond to the recorded daily maximum. The other hour identified as 'high' is the last one recorded and corresponds to a local afternoon maximum and may well have been followed by a large undetected SO_2 value.
77/2/22	75	The maximum monitored value does correspond to the two identified 'high' hourly observations. However, the inversion base remains at below 200 m for this period. The inversion base subsequently rises but the wind speed at the ground and the wind direction shear both decrease.

77/2/9	106	The maximum monitored value corresponds to an identified 'high' hourly observation, it is also the first observation recorded (1 pm) and may well have been preceded by a larger SO_2 value.

4. LAKEVIEW HIGH SO_2 SCENARIOS

It is expected that the scenarios for high SO_2 episodes defined by the two sets of criteria given in Section 3.2 will be site specific. The degree to which these are applicable to other fossil-fired generating station locations is unknown. In order to provide some insight into whether the chosen meteorological variables for Nanticoke are adequate indicators of high pollution episodes at other sites an analysis was made of a small dataset from Lakeview GS.

4.1 Data

The dataset consisted of 167 observations available on 22 days at Lakeview. The observations covered every month of the year except January, February and March. Megawatt (MW) and surface wind speed (WSG) were available for all cases and the wind direction shear (WDS) was available in all but nineteen cases. The height of the inversion base (HIB) was available for 101 cases.

4.2 Criteria for Identifying High SO_2 Days from Meteorological and Load Data

Using the same meteorological variables that were identified for Nanticoke, the high SO_2 cases (days where the maximum reached 120 ppb or more) were inspected with a view to establishing similar criteria for the hourly observations at Lakeview. There were several differences noted between the Nanticoke and Lakeview data, quite apart from the values themselves, the number of observations available for each day at Lakeview was much smaller than for Nanticoke and the observations were generally taken at 40 min rather than at hourly intervals.

No specific values of MW, WSG, WDS could be identified which characterized individual observations on high days. In particular the critical values for WDS found to be associated with high days at Nanticoke were not evident in the Lakeview data. However, there did appear to be basic meteorological states which were associated with high SO_2 episodes, these were typified by either very low WSG, or moderate WSG which at some time of the day exceeded a critical speed. High days were found to occur when either of the following scenarios were present:

a) Daily mean MW > 500 and daily mean WSG < 1 m/s
b) Daily mean MW > 500 and daily mean WSG < 3 m/s with daily maximum WSG > 3.5 m/s.

As a test of the validity of these criteria, all 22 days of Lakeview data were classified as high or non-high days according to the criteria. The results of this exercise are discussed in the following section.

4.3 Predicted and Actual High Days

In Figure 2 the predicted high days, those satisfying either of the two criteria given above, are plotted against the actual daily maximum monitored SO_2 values for the 22 days available from Lakeview. The data is plotted such that where a day was classified as "high" it appears as XXXXX and where it was not, as 00000. It can be seen that the above procedure correctly identifies all of the highest recorded values (those 120 ppb or more). In all, eleven days were classified as high, three non-high days being incorrectly classified.

Fisher's Exact tests (see Section 3.3) performed on the resulting contingency tables give a probability against chance of less than 4 in 100 for the high cases and greater than 1 in 10 for the non-high cases. In other words the observed correct classification rate for high days is superior to that expected from a 'fix' with probability 0.96.

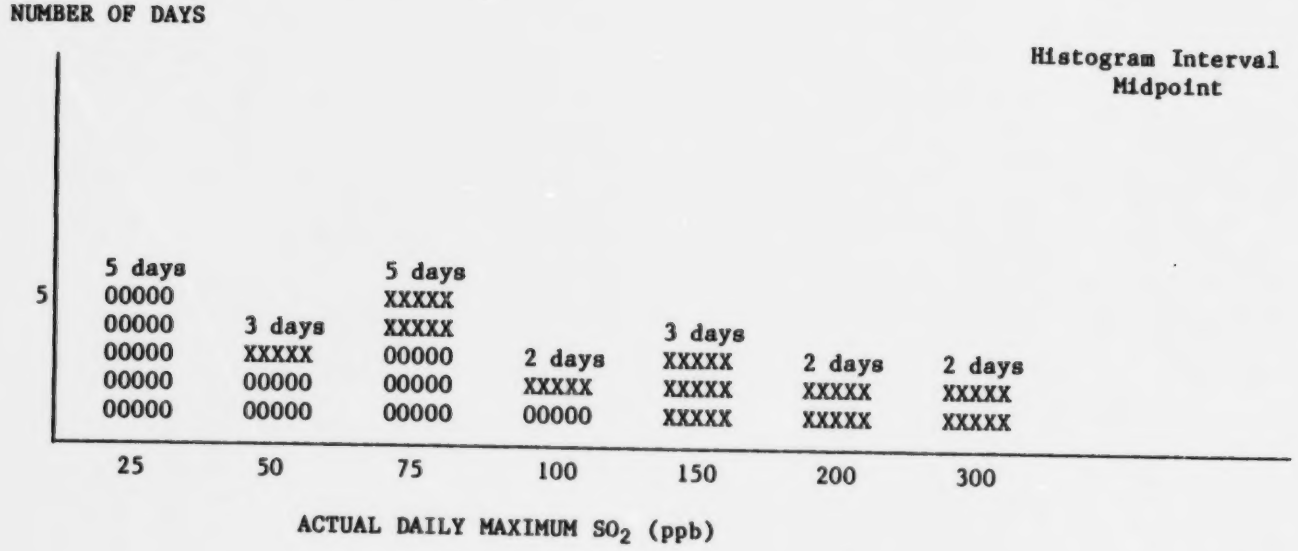
The misclassified cases are described in the following section.

4.4 Misclassified Days

Date	Max-SO_2 (ppb)	Remarks
79/12/11	69	This day was incorrectly classified as "high". The maximum monitor reading was recorded at 6:00 am which is a very unusual time in this dataset. It looks very much like there could have been a missed high afternoon SO_2 value.
80/4/18	71	This day was incorrectly classified as "high". Throughout the period of highest wind speed the values of WDS were too high to allow a high concentration to occur.

FIGURE 2: ACTUAL HIGH DAYS AND PREDICTIONS FOR 22 DAYS AT LAKEVIEW

XXXXX Denotes Predicted High Day
00000 Otherwise



80/4/22	50	This day was incorrectly classified as "high". The WSG values were high and WDS values moderate. It looks very much like there could have been a high SO_2 value missed by the monitors.

5.0 COMPARISON BETWEEN THE FINDINGS FOR NANTICOKE AND LAKEVIEW

An interpretation of the Lakeview results is obviously much more difficult than for Nanticoke because the data available for the former station covers so few days. However, the following points are evident.

1. Where the wind speed at ground level is the sole criterion for high SO_2 episodes, the speed required at Lakeview is generally lower (average > 3 m/s, maximum > 3.5 m/s) than for Nanticoke (minimum > 8 m/s) for a similar time period of at least one to two hours.

2. A Lakeview type scenario was constructed for Nanticoke based only on daily mean and maximum WSG values and MW levels; adjusted for the higher values of all variables at Nanticoke. These were daily mean MW > 1000, daily mean WSG > 8 m/s. daily maximum WSG > 9 m/s. The application of these criteria resulted in poor classifications. The latter failed to identify the two largest SO_2 episodes and performed with only moderate skill in the remaining cases. This showed that the scenarios for high days at Nanticoke and Lakeview are not merely different forms of the same conditions.

3. The importance of wind direction shear as an indicator of activity in the 200-500 m layer appears to be different at Lakeview and Nanticoke. Whereas at Nanticoke certain values of WDS together with moderate WSG invariably indicate activity likely to produce a high SO_2 reading there appears to be no such values at Lakeview. Furthermore, the HIB values in the Lakeview dataset are all in the 200-400 m range which is certainly not the case with the Nanticoke dataset. Thus, the HIB and WDS behaviour at Nanticoke appear to be more complex than at Lakeview, even bearing in mind the small amount of Lakeview data available.

4. Although the results from Nanticoke and Lakeview do not directly correspond they do support the notion, that given moderate generation levels, wind speed at the ground and wind direction shear are the most reliable indicators of the daily maximum SO_2 reading in the vicinity of a generating station.

6.0 SUMMARY

This report describes the basic meteorological variables found to be associated with high SO_2 episodes at Nanticoke GS and gives their critical values. Two scenarios were identified for Nanticoke which accounted for nearly all cases with maximum daily SO_2 values of 120 ppb or greater. These two sets of meteorological scenarios were based on specific values of the following variables:

Wind speed at ground (WSG)
Wind direction shear (WDS)
Total generation (MW)

Because of the very good separation found between high SO_2 cases (120 ppb or greater) and all other cases, the scenarios for non-high SO_2 cases (or situations of high dispersion) can be obtained by reversing the given criteria for high cases.

The analysis of a small set of 22 days data from Lakeview GS confirmed the primary importance of wind speed at ground level as an indicator of SO_2 dispersion. However, the scenarios identified for Lakeview differed from those identified for Nanticoke with respect to the role of wind direction shear.

While the scenarios are site dependent, the approach to scenario identification described in this report is applicable to any similar location.

Approved:

J.G. Cassan
Manager
Operations Research Dept

Submitted:

T.J. Barker.

T.J. Barker
Mathematician - Reliability
and Statistics Section

TJB:km

REFERENCES

1. Barker, T.J. (1982). Database Description and Preliminary Statistical Analysis of SO_2, Meteorological and Station Data at Nanticoke GS (1976-78). Ontario Hydro Research Division Report No 82-56-K.

2. Everitt, B. (1978). Graphical Techniques for Multivariate Data. Heinemann Educational Books, London.

3. Cover, T.M. and Hart, P.E. (1967). Nearest Neighbor Pattern Classification. IEEE Trans. Inform. Theory, Vol IT-13, pp 21-27.

4. Maxwell, A.E. (1961). Analysing Qualitative Data. Chapman and Hall, London.